Universiteit Leiden
Governance and Global Affairs

COMPTEXT CONFERENCE

Text-as-Data Approach to Examining Sustainable Investments

Insights from OECD Financial Institutions

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalezbustamante@fgga.leidenuniv.nl

Natascha van der Zwan
Leiden University
✉ n.a.j.van.der.zwan@fgga.leidenuniv.nl

Presentation delivered at the 7th Annual COMPTEXT Conference
University of Vienna, Austria, April 24-26, 2025

Introduction

Research Focus

Artwork by Leonardo Phoenix model

Asset Owners

Examining insurance companies, pension funds and sovereign wealth funds in OECD countries

Sustainable Investments

Analysing magnitude and nature in relation to institutional variables

Comparative Approach

Identifying impediments and opportunities across different countries

Challenges in Situating Sustainable Finance

Conceptual Challenge

How to place sustainable finance within existing politico-economic typologies

Empirical Challenge

Lack of official SI data and consensus on sustainable investment definition

Greenwashing Concern

Difficulty in distinguishing genuine sustainable practices from greenwashing

Artwork by DALL·E model

Political Economy Approaches to Sustainable Finance

Artwork by DALL·E model

Nature of Sustainable Finance

Studies interrogating green finance and climate finance

Institutional Adoption

Case studies on financial institutions adopting sustainability practices

Policy Domain

Sustainable finance as a new policy area

Global Context

Emergence and diffusion within global financial capitalism

Research Design

Data Collection

1st Stage

- Asset owner directory at www.top1000funds.com
- Proof-of-concept with a reduced scope: top 100 asset owners from Europe. Revealing snapshot of the prevalence of sustainable investment topics

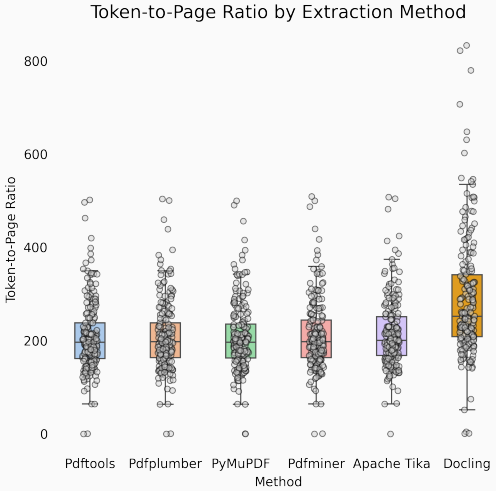
2nd Stage

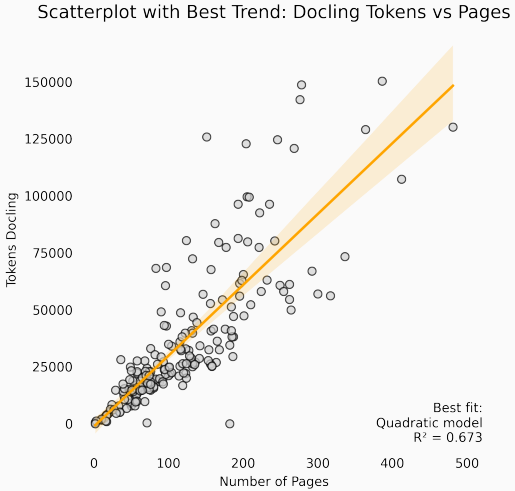
- We expanded the sample to **200 asset owners** worldwide to capture a more global landscape
- We collected, for most cases, the annual reports published in 2023

Parsing Tools

- Docling framework for accurate and comprehensive text extraction
- State-of-the-art AI models for layout recognition

Benchmark of PDFs Parsing Methods





Note. Two observations were excluded from the plots because of their high number of tokens.

Data Overview

32

Countries

Top asset owners from these countries for the moment

24K

Pages

Only annual reports, but we are including specific reports and disclosures

7.6M

Tokens

Tokens for processing and analysis for the moment

Structural Topic Model (STM) Application

1. Machine Translation

- Use GPT-4o/o3-mini to translate 17% of reports into English
- o3-mini allowed an improvement of ∼42% in BLEU/SacreBLEU values

2. Incorporate Metadata

- Include document-level information like original language
- STM directly integrates contextual variables into the modelling process

3. Pre-processing

- Apply standard NLP pipeline for topic modelling
- Converting the text to lowercase, removing punctuation, tokenisation, stopwords list

4. Topic Extraction

- Set STM to extract 15 topics based on coherence and exclusivity
- These measures help ensure that topics are internally coherent and sufficiently distinct

Econometric Strategy

OLS Models

Use per-document topic prevalence (θ-values) from STM as dependent variable

Focus

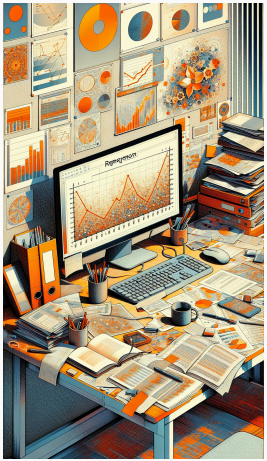
Only on topics relevant to sustainable investment

Independent Variables

AUM, greenhouse gas emissions per capita, carbon intensity of GDP (2021 PPP dollar) and dummy for pension funds

Additional Controls

GDP growth, inflation and regional FE



Artwork by DALL·E model

Preliminary Results

Geographical Distribution and Assets for the 100 European Owners

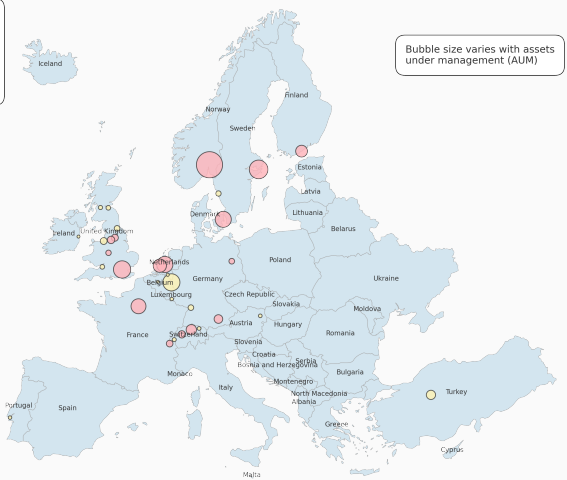
We have created geographical clusters within 50 km based on each owner headquarters location.

This does not imply any relationship between the owners. It is only geographical proximity for visualisation at this stage.

Pastel red bubbles represent clusters (combined assets within 50 km).

Yellow bubbles represent single owners.

Bubble size varies with assets under management (AUM)



Owners and assets under management (AUM)

- Clustered owners
- Single owners

- 137 AUM
- 411 AUM
- 686 AUM
- 961 AUM
- 1,373 AUM

STM Overview

- **15 Topics Identified.** Ranging from mainstream financial practices to sustainability-focused themes
- **National Prevalence.** Some topics show a strong association with specific countries
- **Sustainability Focus.** Three topics stand out in terms of sustainable investments
  - **Climate Emission Reporting.** Focus on CO_2 emissions measurement and reduction targets. Mainly Dutch pension funds and asset managers
  - **Sustainability as Market Practice.** Combining sustainability with existing market practices. Promotion of new sustainable investment products
  - **Climate Leadership.** Emphasis on climate commitments and net zero goals. Mainly Swedish pension funds and insurance companies

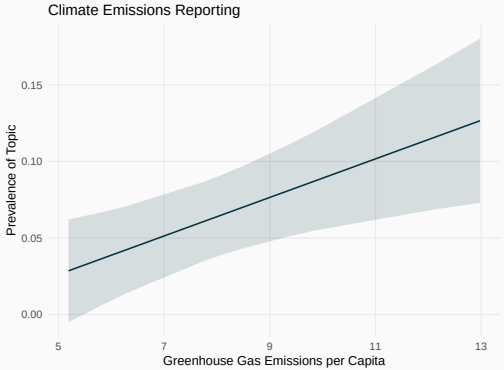
Topics, Labels and Prevalence (100 European Asset Owners)

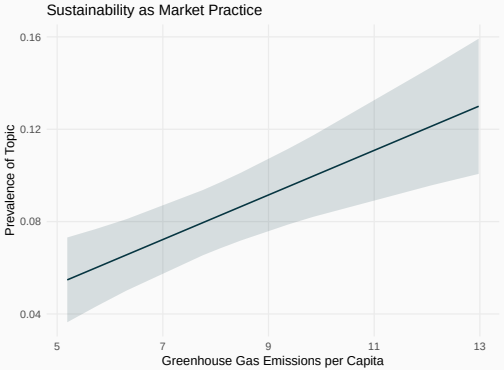
Climate Emission Reporting	Occupational Pension Schemes and Member Befits	Pension Liabilities	Corporate Governance and Diversity	Contribution Monitoring and Regulatory Compliance
Reporting Standards and Strategy Assessments	Portfolio Management and Performance	**Sustainability as Market Practice**	Document Formatting and Linguistic Artefacts	Financial Operations
Group Assets and Credit Risk	**Climate Leadership**	Corporate and Government Bond Holdings	Pension Investment and Fund Returns	Pension Assets

Determinants of Topics Related to Sustainable Investments

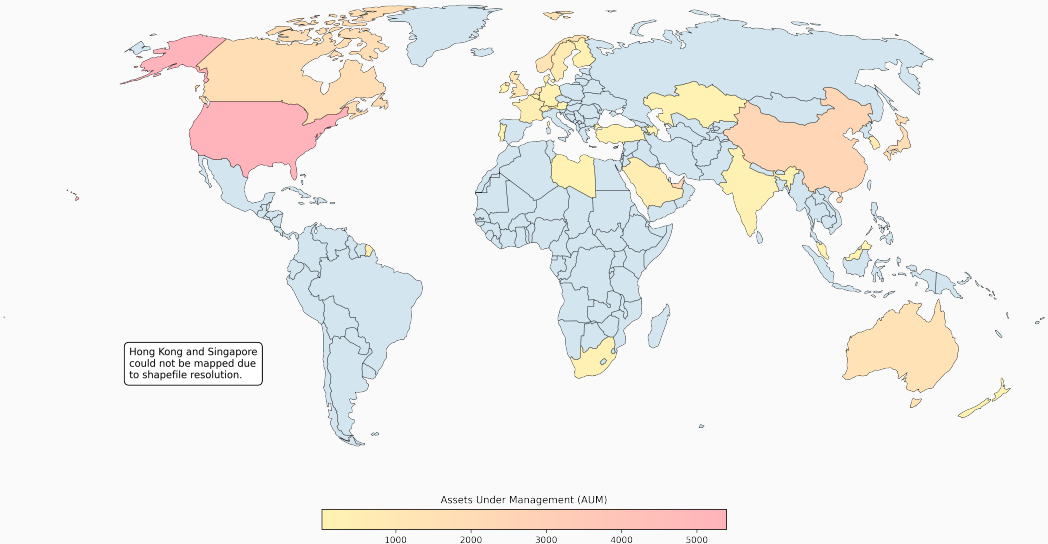
	Climate Emissions Reporting		Sustainability as Market Practice		Climate Leadership	
AUM (1st quartile)	0.002	0.015	0.005	−0.004	0.059**	0.023
	(0.020)	(0.019)	(0.011)	(0.011)	(0.027)	(0.024)
AUM (2nd quartile)	−0.015	−0.003	−0.007	−0.013	0.083***	0.064***
	(0.020)	(0.019)	(0.010)	(0.010)	(0.027)	(0.023)
AUM (3rd quartile)	−0.007	0.001	−0.004	−0.005	0.002	0.005
	(0.020)	(0.019)	(0.011)	(0.010)	(0.027)	(0.023)
Greenhouse gas emissions (per capita)	0.009*	0.013***	0.010***	0.010***	−0.003	−0.001
	(0.005)	(0.004)	(0.002)	(0.002)	(0.006)	(0.006)
Carbon intensity of GDP	0.956***	0.118	−0.485**	−0.348	−0.707	−1.042**
	(0.358)	(0.419)	(0.191)	(0.229)	(0.490)	(0.522)
Constant	−0.083**	−0.111***	0.038**	0.053***	0.100**	0.158***
	(0.032)	(0.032)	(0.017)	(0.018)	(0.044)	(0.040)
GDP growth	No	Yes	No	Yes	No	Yes
Inflation	No	Yes	No	Yes	No	Yes
Estimation method	OLS	OLS	OLS	OLS	OLS	OLS
VIF	1.312	1.508	1.243	1.365	1.216	1.691
N	100	100	100	100	100	100
R^2	0.238	0.337	0.196	0.268	0.178	0.409
Adj. R^2	0.197	0.286	0.153	0.212	0.134	0.364

Effect of Greenhouse Gas Emissions on the Prevalence of SI Topics





Geographical Distribution and Assets for the Top 200 Owners

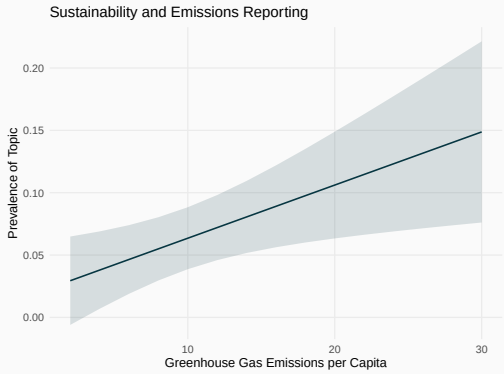


Topics, Labels and Prevalence (200 Top Asset Owners)

Pension Fund Investment and Management	Corporate Structure and Governance	**Sustainability and Emissions Reporting**	Credit and Securities Investment Management	Corporate Governance and ESG Policies
Asset Management and Risk Oversight	Financial Portfolio Management and Risk Control	Public Investment Funds and Fiscal Planning	Corporate Performance and Remuneration Reporting	Pension Contributions and Retirement Benefits
Superannuation Scheme and Corporate Finance	Superannuation Scheme Management and Governance	Insurance and Reinsurance Operations	Global Banking and Investment Development	Liability Valuation and Interest Rate Accounting

Determinants of Topic Related to Sustainable Investments

	Model I	Model II	Model III
AUM (1st quartile)	0.050***	0.041**	0.048***
	(0.018)	(0.017)	(0.016)
AUM (2nd quartile)	0.057***	0.045***	0.054***
	(0.017)	(0.017)	(0.015)
AUM (3rd quartile)	0.040**	0.031**	0.054***
	(0.016)	(0.016)	(0.015)
Greenhouse gas emissions (per capita)	−0.003***	−0.002**	0.004**
	(0.001)	(0.001)	(0.002)
Carbon intensity of GDP	−0.163**	−0.186**	−0.056
	(0.080)	(0.077)	(0.120)
Pension fund (dummy)			−0.009
			(0.011)
Constant	0.088***	0.078***	−0.097
	(0.013)	(0.017)	(0.075)
Region FE	No	No	Yes
GDP growth	No	Yes	Yes
Inflation	No	Yes	Yes
Estimation method	OLS	OLS	OLS
VIF	1.176	1.311	1.695
N	200	200	200
R^2	0.149	0.237	0.410
Adj. R^2	0.128	0.209	0.372



Takeaways

Takeaways

☑ **Regional Context.** Sustainability topics shaped by regional and institutional factors.

☑ **Emission Impact.** Higher emissions linked to more emphasis on market-based approaches and CO_2 tracking.

☑ **European Propensity.** Top European asset owners show a higher variety of sustainability topics, which dilutes when expanding the sample worldwide.

- **Limitations.** (1) Text-based measures are susceptible to biases like greenwashing or cheap talk in annual reports. (2) Differences in reporting frameworks across firms and countries.
- **Next Steps.** (1) Move beyond the top 200 owners. (2) Additional variables such as regulation indices or network memberships.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub